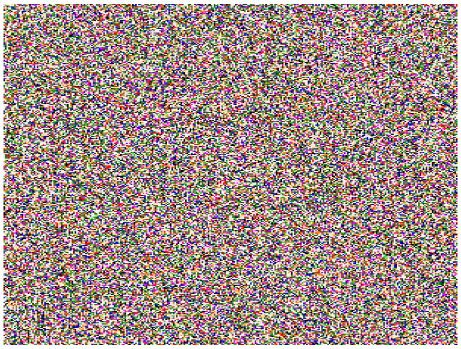
Data Compression - Lecture 1 (handout notes)

[Part of seminar series on "Data Compression" created by Elias Machairas, eliasmach@di.uoa.gr]

Data Compression is split into two major subfields: Lossless Data Compression and Lossy Data Compression.
The goal of Data Compression is to **reduce the size of any file** (text, image, video, audio, etc) **as much as possible without losing** any (in case of Lossless) or much (in case of Lossy) **information**. That is, later **when** the compressed file gets **decompressed we are able to retrieve** all (Lossless) or most (Lossy) of **the original data**. Let's first consider the following question:

1. Which of the following two images (of same size) contains the most information ?



(a)

(b)

Solution:

It turns out, as you would expect, that not all data contain the same amount of information and this is directly linked to how much we can compress them. But what is information

? We may feel comfortable with the notion of information in our everyday lives but it is difficult to pin it down exactly. In the following sections, we will give the precise established mathematical definition of information and link it to our usual notion of information from everyday life. However, **the mathematical definition of information will probably be the opposite of what you would expect and answered in question 1**, and we will see the reasons why it is useful to define and think of it that way. In short, we will define information such that the following natural facts are true: 1) less information leads to less size and 2) more information leads to more size. So let's dive right in and turn our vague notion of information into a mathematical object that will guide all of our endeavors in compressing data. One more thing to be done before we begin is to give credits to Claude Shannon (1916 - 2001), the scientist who invented all of the mathematical ideas that will concern us in this course which are collectively called 'Information Theory' and are the mathematical underpinnings of data compression.

As will be the general format of this course, we will explore new ideas by answering a series of numbered questions/discussions (each followed by its own boxed area). You've already encountered the first question above (question 1). Now let's continue to discussion number 2, i.e. the mathematical definition of information.

2. Mathematical definition of information

Solution:

3. Adding the law of large numbers to our toolbox

Solution:

4. Is data compression possible philosophically ? Put differently, how can we make something smaller and not lose any information at all ? We solve this mystery using probability theory.

Solution:

Exercise 1: Sampling from a distribution and storing to file.
Let X be a random variable with the following distribution: $P_X(x) = (\frac{1}{4}, \frac{1}{8}, \frac{1}{8}, \frac{1}{2})$.
a) Compute its entropy, H(X).
b) If you sample from this distribution 10000 times and you want to compress (using your own breakthrough compression algorithm!) and store all the samples inside a file, what is the minimum file size you should aim for ?

Solution:

Exercise 2: Is Entropy a bad (invalid) metric for real data?
Out of boredom, you typed the following string: ”12345678123456781234567812345678” into a file and now you want to compress it.
a) Compute the entropy of the input string.
b) Can you beat the entropy (i.e. compress the file using less bits) ? Explain when (for which inputs) the entropy bound is valid (useful) and when it isn't.
c) What is the general methodology to mitigate this situation when we develop real compression algorithms for real data ?

Solution: